CS3236 Introduction to Information Theory

AY2022/23 Semester 2 · Prepared by Tian Xiao *@snoidetx*

1 Information Measures

Information of an Event: If event A occurs with probability p, then we have

$$\text{Info}(A) = \psi(p) = \log_b \frac{1}{p}.$$

- When $b = 2$, information is measured in bits.
- Axiomatization of $\psi(p)$:
  - ▷ `Non-Negativity`: $\psi(p) > 0$;
  - ▷ `Zero for Definite Events`: $\psi(1) = 1$;
  - ▷ `Monotonicity`: $p \le p' \Rightarrow \psi(p) \ge \psi(p')$;
  - ▷ `Continuity`: $\psi(p)$ is continuous in p;
  - ▷ `Additivity under Independence`: $\psi(p_1 p_2) = \psi(p_1) + \psi(p_2)$.

Shannon Entropy: Let X be a discrete random variable with probability mass function P_X. The Shannon entropy of X is the average information we learn from observing $X = x$ (note: $0 \log_2 \frac{1}{0} = 0$):

$$H(X) = \mathbb{E}_{X \sim P_X}[\psi(X = x)] = \sum_x P_X(x) \log_2 \frac{1}{P_X(x)}.$$

- Joint entropy:
$$H(X, Y) = \mathbb{E}_{(X,Y)\sim P(X,Y)}[\psi(X = x, Y = y)] = \sum_{x,y} P_{XY}(x, y) \log_2 \frac{1}{P_{XY}(x, y)}.$$
- Conditional entropy:
$$H(Y|X) = \mathbb{E}_{(X,Y)\sim P(X,Y)}[\psi(Y = y|X = x)] = \sum_{x,y} P_{XY}(x, y) \log_2 \frac{1}{P_{Y|X}(y|x)} = \sum_x P_X(x) H(Y|X = x).$$
- Entropy measures information or uncertainty in X.
  - ▷ Binary source: $H(X) = H_2(p) = p \log_2 \frac{1}{p} + (1 - p) \log_2 \frac{1}{1-p}$;
  - ▷ Uniform source: $H(X) = \log_2 |\mathcal{X}|$.
- Axiomatization of $\Psi(\mathbf{p})$: Suppose that X is a discrete random variable taking N values with probabilities $\mathbf{p} = \{p_1, \cdots, p_N\}$. Consider an information measure $\Psi(\mathbf{p}) = \Psi(p_1, \cdots, p_n)$:
  - ▷ `Continuity`: $\Psi(\mathbf{p})$ is continuous as a function of $\mathbf{p}$;
  - ▷ `Uniform Case`: If $\forall i \left[p_i = \frac{1}{N}\right]$, then $\Psi(\mathbf{p})$ is increasing in N;
  - ▷ `Successive Decisions`:
$$\Psi(p_1, \cdots, p_N) = \Psi(p_1 + p_2, p_3, \cdots, p_N) + (p_1 + p_2)\Psi\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right).$$
- Properties of entropy:
  - ▷ `Non-Negativity`: $H(X) \ge 0$;
  - ▷ `Upper Bound`: $H(X) \le \log_2 |X|$;
  - ▷ `Chain Rule` (2 var): $H(X, Y) = H(X) + H(Y|X)$;
  - ▷ `Chain Rule` (n var):
$$H(X_1, \cdots, X_n) = \sum_{i=1}^{n} H(X_i|X_1, \cdots, X_{i-1});$$
  - ▷ `Conditioning Reduces Entropy`: $H(X|Y) \le H(X)$ with equality if and only if X and Y are independent;
  - ▷ `Sub-Additivity`: $H(X_1, \cdots, X_n) \le \sum_{i=1}^{n} H(X_i)$.

KL Divergence:

$$D(P||Q) = \mathbb{E}_{X\sim P}\left[\log_2 \frac{P(x)}{Q(x)}\right] = \sum_x P(x) \log_2 \frac{P(x)}{Q(x)}.$$

- $D(P||Q) \ge 0$ with equality if and only if $P = Q$.

Mutual Information: Information between random variables:

$$I(X; Y) = H(Y) - H(Y|X).$$

- Terminologies:
  - ▷ $H(Y)$: Prior uncertainty in Y;
  - ▷ $H(Y|X)$: Remaining uncertainty in Y after observing X;
  - ▷ $I(X; Y)$: Information we learn about Y after observing X.
- Joint mutual information:
$$I(X_1, X_2; Y_1, Y_2) = H(Y_1.Y_2) - H(Y_1, Y_2|X_1, X_2).$$
- Conditional mutual information:
$$I(X; Y|Z) = H(Y|Z) - H(Y|X, Z).$$
- Properties of mutual information:
  - ▷ `Alternative Forms`:
$$I(X; Y) = D(P_{XY}||P(X) \times P(Y)) = \sum_{x,y} P_{XY}(x, y) \log_2 \frac{P_{XY}(x, y)}{P_X(x)P_Y(y)} = \sum_{x,y} P_{XY}(x, y) \log_2 \frac{P_{Y|X}(y|x)}{P_Y(y)};$$
  - ▷ `Symmetry`: $I(X; Y) = I(Y; X) = H(X) + H(Y) - H(X, Y)$;
  - ▷ `Non-Negativity`: $I(X; Y) \ge 0$ with equality if and only if X and Y are independent;
  - ▷ `Upper Bounds`: $I(X; Y) \le H(X)$; $I(X; Y) \le H(Y)$.
  - ▷ `Chain Rule`: $I(X_1, \cdots, X_n|Y) = \sum_{i=1}^{n} I(X_i; Y|X_1, \cdots, X_{i-1})$;
  - ▷ `Data Processing Inequality`: If X and Z are conditionally independent given Y, then $I(X; Z) \le I(X; Y)$;
  - ▷ `Partial Sub-Additivity`: If $Y_1, \cdots, Y_n$ are conditionally independent given $X_1, \cdots, X_n$, and Y_i depends on $X_1, \cdots, X_n$ only through X_i, then
$$I(X_1, \cdots, X_n; Y_1, \cdots, Y_n) \le \sum_{i=1}^{n} I(X_i; Y_i).$$

2 Symbol-Wise Source Coding

Symbol-Wise Coding: Symbol-wise source coding maps each $x \in \mathcal{X}$ to some binary sequence $C(x)$. The length of this sequence is denoted by $\ell(x)$. The average length of a code $C(\cdot)$ is given by

$$L(C) = \sum_{x\in\mathcal{X}} P_X(x)\ell(x).$$

- `Non-Singular`: $x \ne x' \Rightarrow C(x) \ne C(x')$;
- `Uniquely Decodable`: A code $C(\cdot)$ is said to be uniquely decodable if no two sequences of symbols in $\mathcal{X}$ are coded to the same concatenated binary sequence;
- `Prefix-Free`: A code $C(\cdot)$ is said to be prefix-free if no codeword is a prefix of any other.

Kraft's Inequality: Any prefix-free code $C(\cdot)$ that maps each $x \in \mathcal{X}$ to a codeword of length $\ell(x)$ must satisfy

$$\sum_{x\in\mathcal{X}} 2^{-\ell(x)} \le 1.$$

- `Existence`: If a set of integers $\{\ell(x)\}_{x\in\mathcal{X}}$ satisfies $\sum_{x\in\mathcal{X}} 2^{-\ell(x)} \le 1$, then it is possible to construct a prefix-free code that maps each $x \in \mathcal{X}$ to a codeword of length $\ell(x)$.

Entropy Bound: For $X \sim P_X$ and any prefix-free code $C(\cdot)$, the expected length satisfies

$$L(C) \ge H(X),$$

with equality if and only if $P_X(x) = 2^{-\ell(x)}$ for all $x \in \mathcal{X}$.

Shannon-Fano Code: $\ell(x) = \left\lceil \log_2 \frac{1}{P_X(x)} \right\rceil$.

- $H(X) \le L(C) < H(X) + 1$.
- If the true distribution is P_X but the lengths are chosen according to Q_X, then the Shannon-Fano code satisfies
$$H(X) + D(P_X||Q_X) \le L(C) \le H(X) + D(P_X||Q_X) + 1.$$

Huffman Code: Construct a tree as follows:

1. List the symbols of $\mathcal{X}$ from highest probability to lowest.
2. Draw a branch connecting the two symbols with the lowest probability, and label the merged point with the sum of the two associated probabilities.
3. Repeat the first two steps until everything has merged to a single point with total probability 1.

- No uniquely decodable symbol code can achieve a smaller average length $L(C)$ than the Huffman code.
- $H(X) \le L(C) < H(X) + 1$.

3 Block Source Coding

Problem Description:

- Source: $\mathbf{X} = (X_1, X_2, \cdots, X_n)$.
  - ▷ `Discrete`: The alphabet $\mathcal{X}$ is finite.
  - ▷ `Memoryless`: $P_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} P_X(x_i)$ (i.i.d.).
- Encoder: Received source $\mathbf{X} \to$ message $m = f(\mathbf{X}) \in \{1, \cdots, M\}$.
- Decoder: Message $m \to$ estimate $\hat{\mathbf{X}} = g(m)$.
- Error probability: $P_e = \mathbb{P}[\hat{\mathbf{X}} \ne \mathbf{X}]$.
- Rate: Number of bits per source symbol: $R = \frac{1}{n} \log_2 M$.

Fix-Length Source Coding Theorem:

- `Achievability`: If $R > H(X)$, then for any $\epsilon > 0$, there exists a sufficiently large block length n and a source code (i.e. encoder and decoder) of rate R such that $P_e < \epsilon$.
- `Converse`: If $R < H(X)$, then there exists $\epsilon > 0$ such that every code of rate R has $P_e < \epsilon$, regardless of the code length.

Typical Set:

$$\mathcal{T}_n(\epsilon) = \left\{\mathbf{x} \in \mathcal{X}^n : 2^{-n(H(X)+\epsilon)} \leq P_{\mathbf{X}}(\mathbf{x}) \leq 2^{-n(H(X)-\epsilon)}\right\}.$$

- Equivalent definition:

$$H(X) - \epsilon \leq \frac{1}{n}\sum_{i=1}^{n} \log_2 \frac{1}{P_X(x_i)} \leq H(X) + \epsilon.$$

- Properties:
  - ▷ `High Probability`: $\mathbb{P}[\mathbf{X} \in \mathcal{T}_n(\epsilon)] \to 1$ as $n \to \infty$.
  - ▷ `Cardinality Upper Bound`: $|\mathcal{T}_n(\epsilon)| \leq 2^{n(H(X)+\epsilon)}$.
  - ▷ `Cardinality Lower Bound`: $|\mathcal{T}_n(\epsilon)| \geq (1 - o(1))2^{n(H(X)-\epsilon)}$, where $o(1)$ represents a term that vanishes as $n \to \infty$.
  - ▷ `Asymptotic Equipartition`: With `High Probability`, a randomly drawn i.i.d. sequence $\mathbf{X}$ will be one of roughly $2^{nH(X)}$ sequences, each of which has probability roughly $2^{-nH(X)}$.

Fano's Inequality: $H(X|\hat{X}) \leq H_2(P_e) + P_e \log_2(|\mathcal{X}| - 1)$.

4 Channel Coding

Problem Description:

- Channel: The medium over which we transit information.
  - ▷ `Discrete`: $\mathcal{X}$ and $\mathcal{Y}$ are finite.
  - ▷ `Memoryless`: Outputs are conditionally independent, i.e.

$$P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{n} P_{Y|X}(y_i|x_i).$$

- Encoder: Message $m \to$ codeword $\mathbf{x}^{(m)} = \left(x_1^{(m)}, \cdots, x_n^{(m)}\right)$.
  - ▷ Codebook $\mathcal{C}$: Collection of codewords $\left\{\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(M)}\right\}$.
- Decoder: Received codeword $\mathbf{y} = (y_1, \cdots, y_n) \to$ estimate $\hat{m}$.
- Error probability: $P_e = \mathbb{P}[\hat{m} \neq m]$.
- Rate: Number of bits per channel use ($R = \frac{1}{n}\log_2 M$).
  - ▷ $M = 2^{nR}$.

Channel Capacity: The channel capacity C is defined to be the maximum of all rates R such that for any target error probability $\epsilon > 0$, there exists a block length n and codebook $\mathcal{C} = \left\{\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(M)}\right\}$ with $M = 2^{nR}$ codewords such that $P_e < \epsilon$.

- Channel Coding Theorem The capacity of a discrete memoryless channel $P_{Y|X}$ is

$$C = \max_{P_X} I(X;Y).$$

  - ▷ `Achievability`: For any $R < C$, there exists a code of rate at least R with arbitrarily small error probability (via random coding).
  - ▷ `Converse`: For any $R > C$, any code of rate at least R cannot have arbitrarily small error probability (via Fano's Inequality).
- Capacity achieving input distribution: Any input distribution P_X maximizing the mutual information above for a given channel $P_{Y|X}$.

Jointly Typical Set:

$$\mathcal{T}_n(\epsilon) = \left\{ (\mathbf{x},\mathbf{y}) : \begin{array}{r} 2^{-n(H(X)+\epsilon)} \leq P_{\mathbf{X}}(\mathbf{x}) \leq 2^{-n(H(X)-\epsilon)} \\ 2^{-n(H(Y)+\epsilon)} \leq P_{\mathbf{Y}}(\mathbf{y}) \leq 2^{-n(H(Y)-\epsilon)} \\ 2^{-n(H(X,Y)+\epsilon)} \leq P_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y}) \leq 2^{-n(H(X,Y)-\epsilon)} \end{array} \right\}.$$

- `High Probability`: $\mathbb{P}[(\mathbf{X},\mathbf{Y}) \in \mathcal{T}_n(\epsilon)] \to 1$ as $n \to \infty$.
- `Cardinality Upper Bound`: $|\mathcal{T}_n(\epsilon)| \leq 2^{n(H(X,Y)+\epsilon)}$.
- If $(\mathbf{X}', \mathbf{Y}') \sim P_{\mathbf{X}}(\mathbf{x}')P_{\mathbf{Y}}(\mathbf{y}')$ are independent copies of $(\mathbf{X},\mathbf{Y})$, then the probability of joint typicality is

$$\mathbb{P}[(\mathbf{X}', \mathbf{Y}') \in \mathcal{T}_n(\epsilon)] \leq 2^{-n(I(X;Y)-3\epsilon)}.$$

5 Continuous Alphabet Channels

Differential Entropy: For a continuous random variable X,

$$h(X) = \mathbb{E}_{f_X}\left[\log_2 \frac{1}{f_X(X)}\right] = \int_{\mathbb{R}} f_X(x)\log_2 \frac{1}{f_X(x)}\,\mathrm{d}x.$$

- Joint entropy:

$$h(X,Y) = \mathbb{E}_{f_{XY}}\left[\log_2 \frac{1}{f_{XY}(X,Y)}\right].$$

- Conditional entropy:

$$h(Y|X) = \mathbb{E}_{f_{XY}}\left[\log_2 \frac{1}{f_{Y|X}(Y|X)}\right] = \int_{\mathbb{R}} f_X(x)h(Y|X=x)\,\mathrm{d}x.$$

- Properties of differential entropy:
  - ▷ `Chain Rule`: $h(X_1, \cdots, X_n) = \sum_{i=1}^{n} h(X_i|X_1, \cdots, X_{i-1})$.
  - ▷ `Conditioning Reduces Entropy`: $h(X|Y) \leq h(X)$.
  - ▷ `Sub-Addivity`: $h(X_1, \cdots, X_n) \leq \sum_{i=1}^{n} h(X_i)$.
  - ▷ $h(X) = h(X + c)$ for any constant c.
  - ▷ `Non-Negativity` and `Invariance Under 1-1 Transformation` no longer holds.
- Examples:
  - ▷ Uniform source $X \sim \mathrm{Uniform}(a,b)$: $h(X) = \log_2(b-a)$.
  - ▷ Gaussian source $X \sim N(\mu, \sigma^2)$: $h(X) = \frac{1}{2}\log_2(2\pi e\sigma^2)$.

KL Divergence: $D(f\|g) = \int_{\mathbb{R}} f(x)\log_2 \frac{f(x)}{g(x)}\,\mathrm{d}x$.

Mutual Information:

$$\begin{aligned} I(X;Y) = D(f_{XY}\|f_X \times f_y) &= \mathbb{E}_{f_{XY}}\left[\log_2 \frac{f_{XY}(x,y)}{f_X(x)f_Y(y)}\right]. \\ &= h(Y) - h(Y|X) = h(X) - h(X|Y) \end{aligned}$$

- All key properties still hold, including `Non-Negativity`.
- For invertible functions ϕ and ψ, $I(X;Y) = I(\phi(X);\psi(Y))$.

Gaussian Random Variables: $X \sim N(\mu, \sigma^2)$.

- `Maximum Entropy Property`: For any random variable X with p.d.f. f_X and variance $\mathrm{Var}[X]$, $h(X) \leq \frac{1}{2}\log_2(2\pi e\mathrm{Var}[X])$.

Gaussian Channel:

- Channel capacity: $C(P) = \max_{f_X:\mathbb{E}_{f_X}[X^2]\leq P} I(X;Y)$.
  - ▷ P is the power constraint.
  - ▷ For the Additive White Gaussian Noise (AWGN) channel with power constraint P and noise variance σ^2, the channel capacity is

$$C(P) = \frac{1}{2}\log_2\left(1 + \frac{P}{\sigma^2}\right).$$

6 Practical Channel Codes

Linear Code: Any code with parity checks is a linear code.

- Types of linear code $\mathbf{u} \to \mathbf{x}$:
  - ▷ Systematic parity-check code: The first k out of n bits of $\mathbf{x}$ are always precisely the original k bits, and the remaining $n-k$ bits are parity checks.
  - ▷ General parity-check code: All n codeword bits may be arbitrarily parity checks.
- Generator matrix: $\mathbf{x} = \mathbf{uG}$, $\mathbf{G}$ is the generator matrix.
- `Linearity`: $\mathbf{x} \oplus \mathbf{x}' = (\mathbf{u} + \mathbf{u}')\mathbf{G}$.
- Parity-check matrix: $\mathbf{xH} = \mathbf{0} \Leftrightarrow \mathbf{x}$ is valid.
  - ▷ For systematic codes, $\mathbf{G} = \begin{bmatrix}\mathbf{I}_k & \mathbf{P}\end{bmatrix} \Rightarrow \mathbf{H} = \begin{bmatrix}\mathbf{P} \\ \mathbf{I}_{n-k}\end{bmatrix}$.

Distance Properties:

- Hamming distance: The Hamming distance between two vectors $\mathbf{x}$ and $\mathbf{x}'$ is the number of positions in which they differ:

$$d_H(\mathbf{x},\mathbf{x}') = \sum_{i=1}^{n} \mathbb{I}[x_i \neq x_i'].$$

- Minimum distance: The minimum distance of a codebook $\mathcal{C}$ of length-n codewords is

$$d_{\min = } \min_{\mathbf{x}\neq\mathbf{x}'\in\mathcal{C}} d_H(\mathbf{x},\mathbf{x}').$$

.

  - ▷ If minimum distance is $d_{\min}$, then it is possible to correct up to $d_{\min -1}$ erasures and $\frac{d_{\min -1}}{2}$ bit flips.
- Weight: $w(\mathbf{x}) = \sum_{i=1}^{n} \mathbb{I}[x_i = 1]$.
  - ▷ For linear codes, minimum distances equal minimum weights.

Minimum Distance Decoding:

- Maximum-likelihood decoder: For any channel $P_{\mathbf{Y}|\mathbf{X}}$ and any codebook $\left\{\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(M)}\right\}$, the decoding rule that minimizes the error probabiltiy P_e is the maximum-likelihood (ML) decoder:

$$\hat{m} = \arg\max_{j=1,\cdots,M} P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}^{(j)}).$$

  - ▷ For a linear code, if the syndrome is $\mathbf{S} = \mathbf{yH} = \mathbf{zH}$, then the minimum-distance codeword to $\mathbf{y}$ can be obtained by finding

$$\hat{\mathbf{z}} = \arg\min_{\tilde{\mathbf{z}}:\tilde{\mathbf{z}}\mathbf{H}=\mathbf{S}} w(\tilde{\mathbf{z}}),$$

  then computing $\hat{\mathbf{x}} = \mathbf{y} \oplus \hat{\mathbf{z}}$.